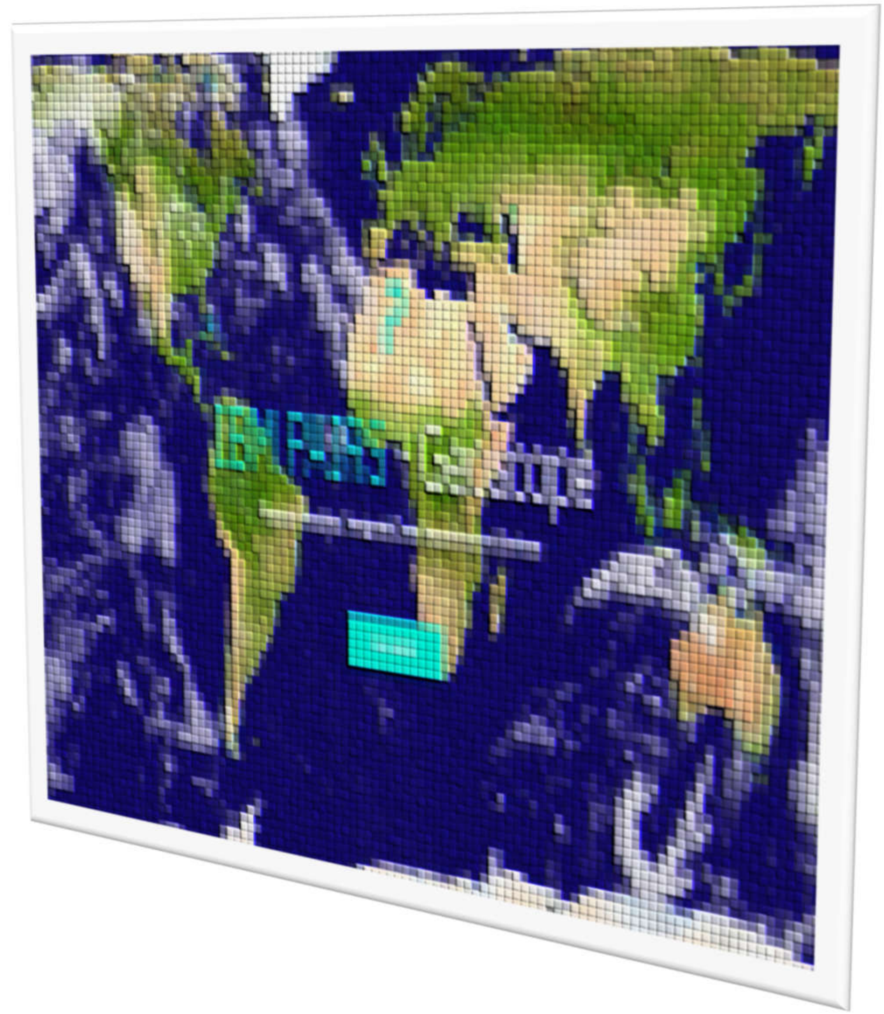
*IBM PAIRS**
- Big Geospatio-temporal Data and Analytics as-a-Service



Siyuan Lu[†], et al.

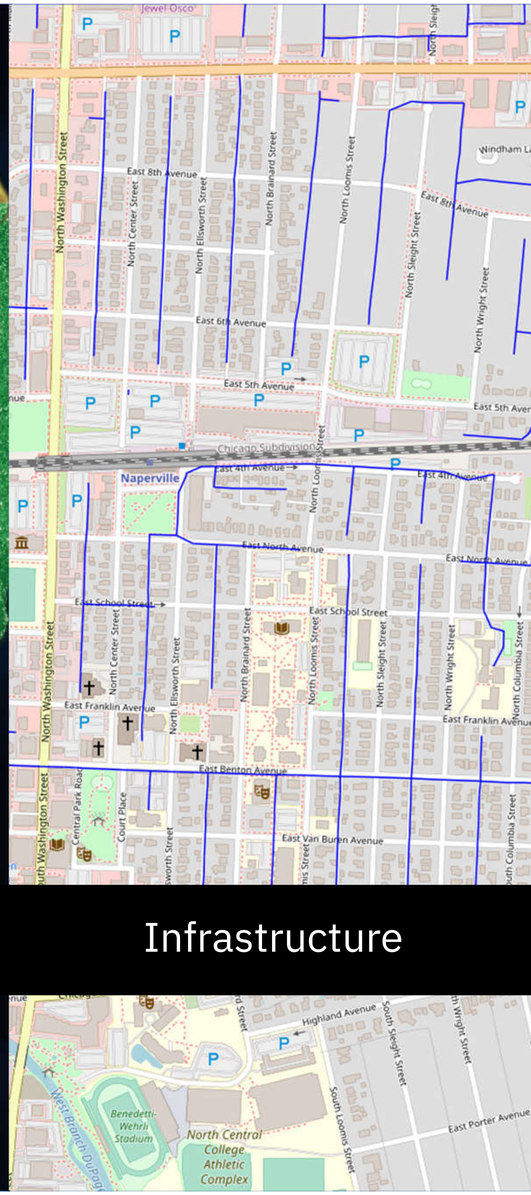
Manager, Principal Research Staff Member

IBM TJ Watson Research Center

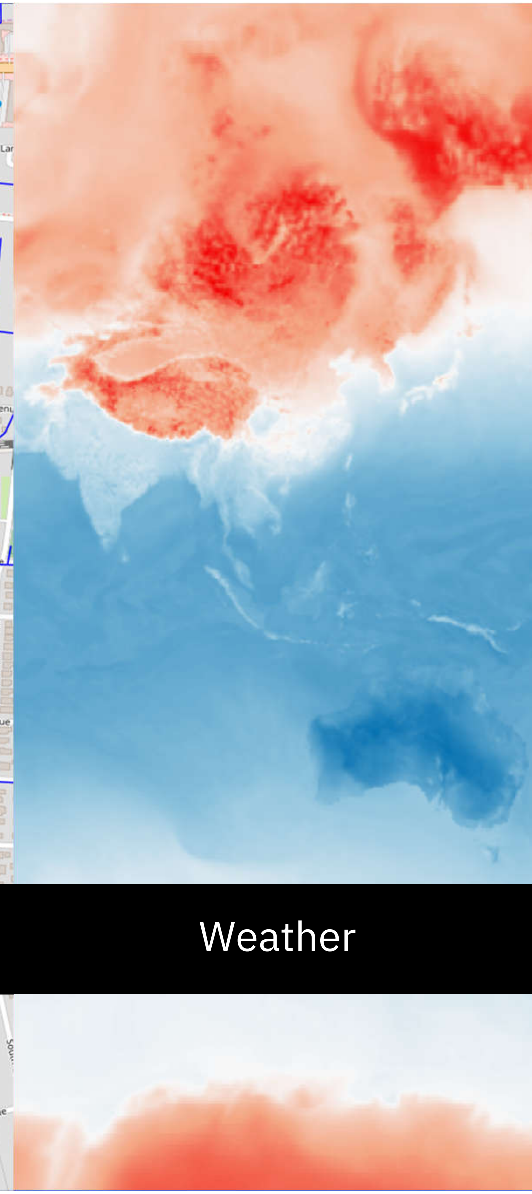
* Physical Analytics Integrated Repository and Services

[†] email: lus@us.ibm.com

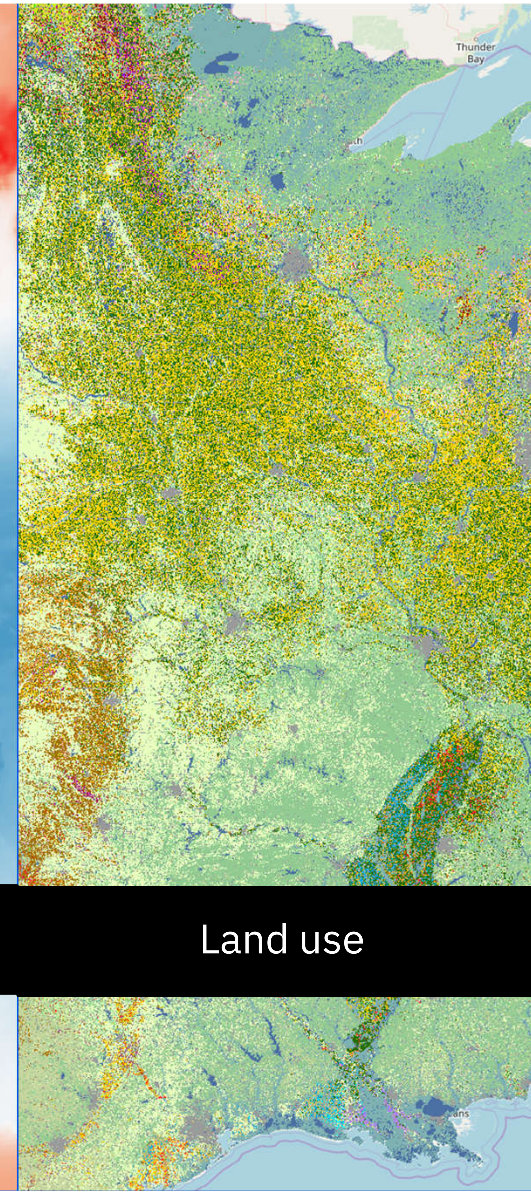
Satellite/drone imagery



Infrastructure



Weather



Land use

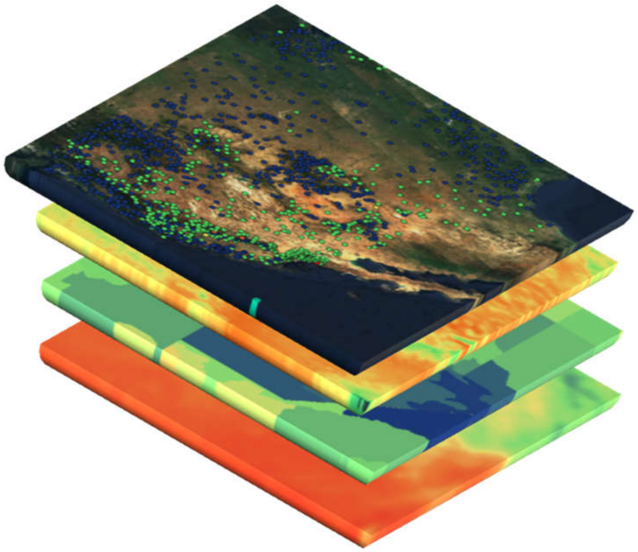
IBM PAIRS Geoscope to enable scalable geospatial-temporal data analytics

(i) Harmonization - masking the complexity of metadata.

(ii) Complex query - searching of data at a "pixel level".

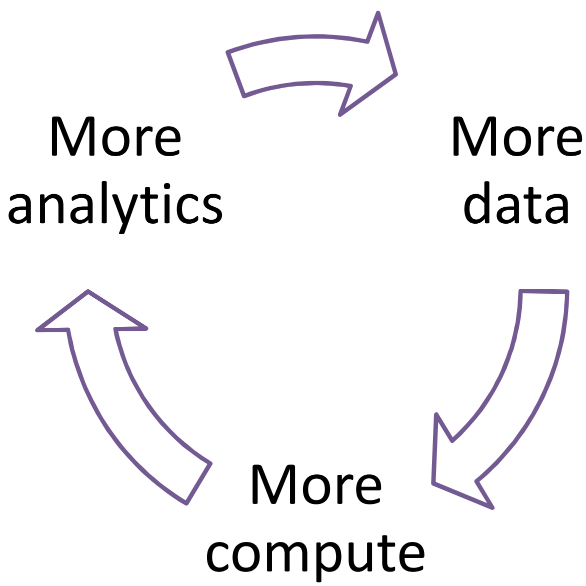
(iii) Scalability - handling petabytes of data

(iv) In-data" computation – avoiding data movement through the internet



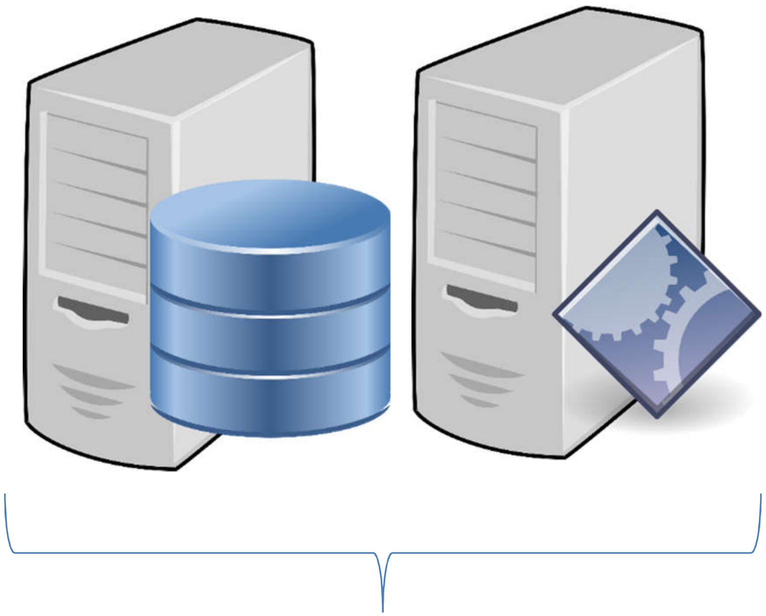
Data gravity - Big data attracts

- Data is too big to be moved
- Context of data is to be exploit



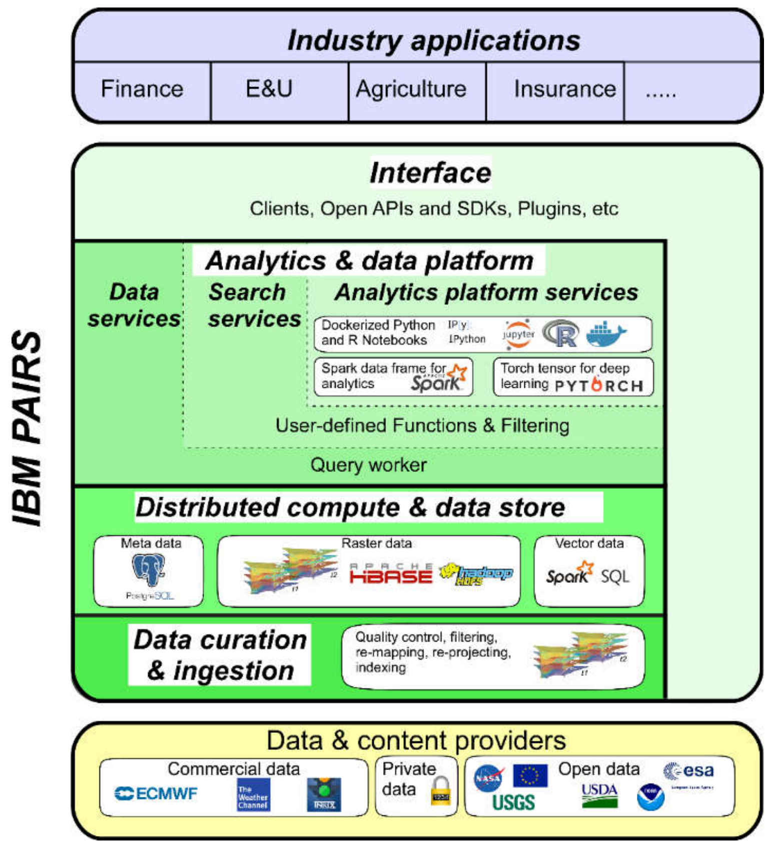
Platform /Data base
85 % of analytics

Application
15 % of analytics



Data and analytics platform

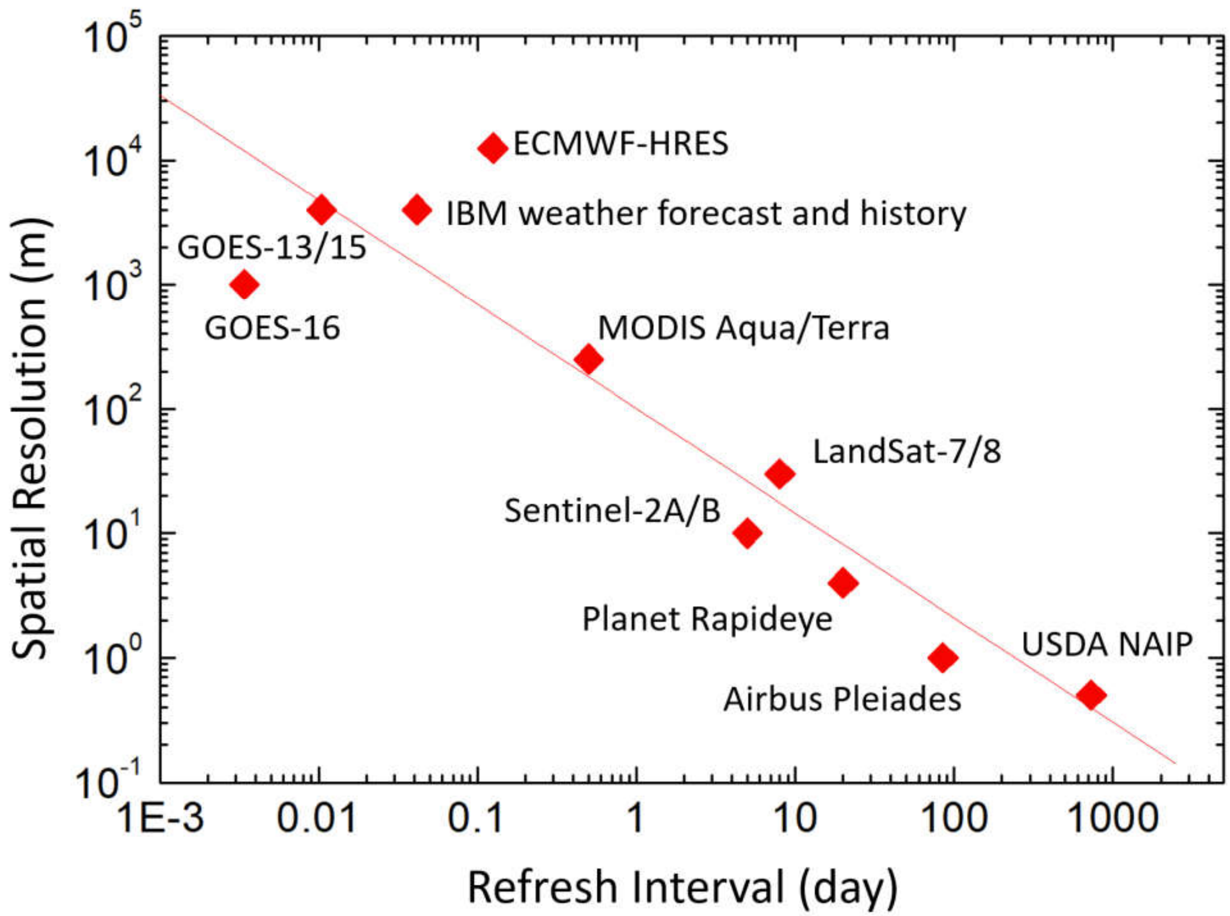
What is PAIRS? A scalable big geospatial-temporal data and analytics platform with cross-industry applications



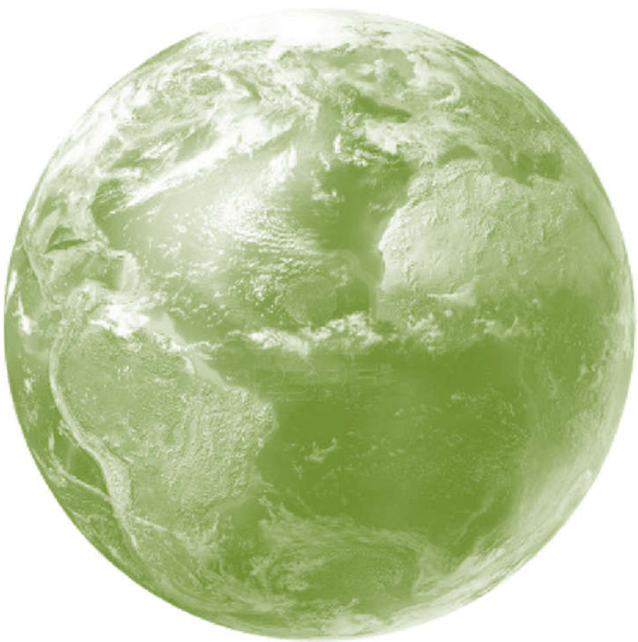
- ✓ **Reproject incoming data into a set of nested global grids.**
- ✓ **Main data store: scalable, low latency, multi-level key-value store (Hbase)**
- ✓ Ingests, curates 20TB+/day of new data
- ✓ 4 PBs of industry relevant data
- ✓ Accessible through APIs/SDKs/GUI for querying and user-enabled data uploads
- ✓ **Query and analytics parallelized using Map-Reduce and SPARK.**

PAIRS Design Decision

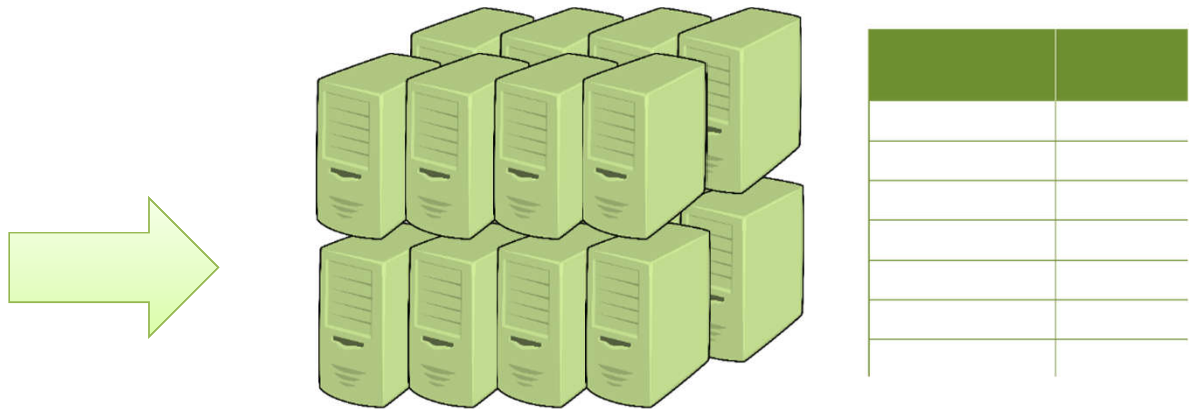
Design Decision
Use a fixed coordinate reference system and a fixed set of nested grids.
Employ a multi-aspect row key to encode spatial and temporal information.
Supercells (32x32 pixels) as the value of the key-value store.
Temporal hash in the HBase row keys.
Spatial hash in the HBase row keys.
Construct a set of coarse-grained overview layers for each data layers.



Challenges lies in translating multi-dimensional data to 1D key-value store to optimally support queries

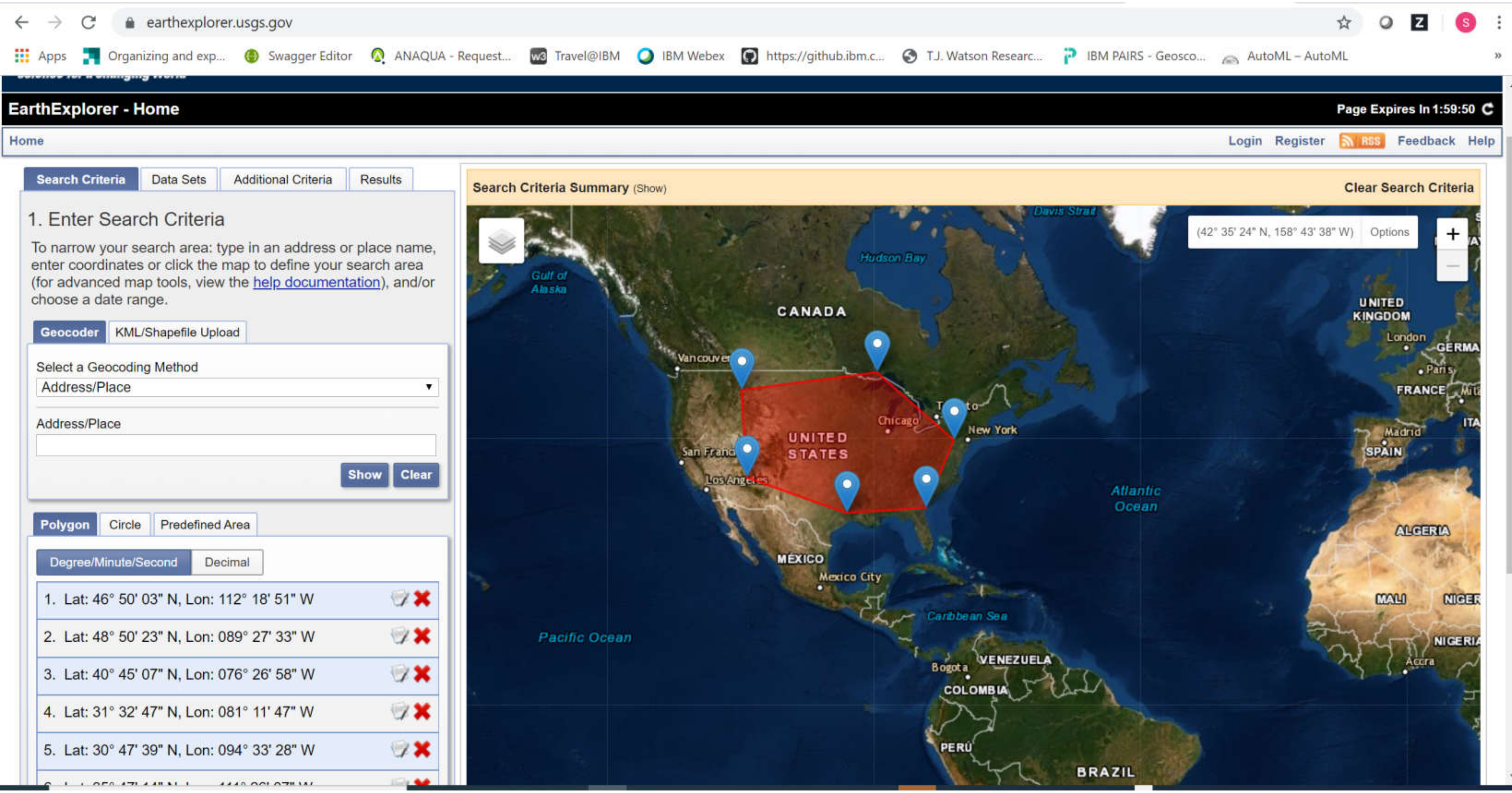


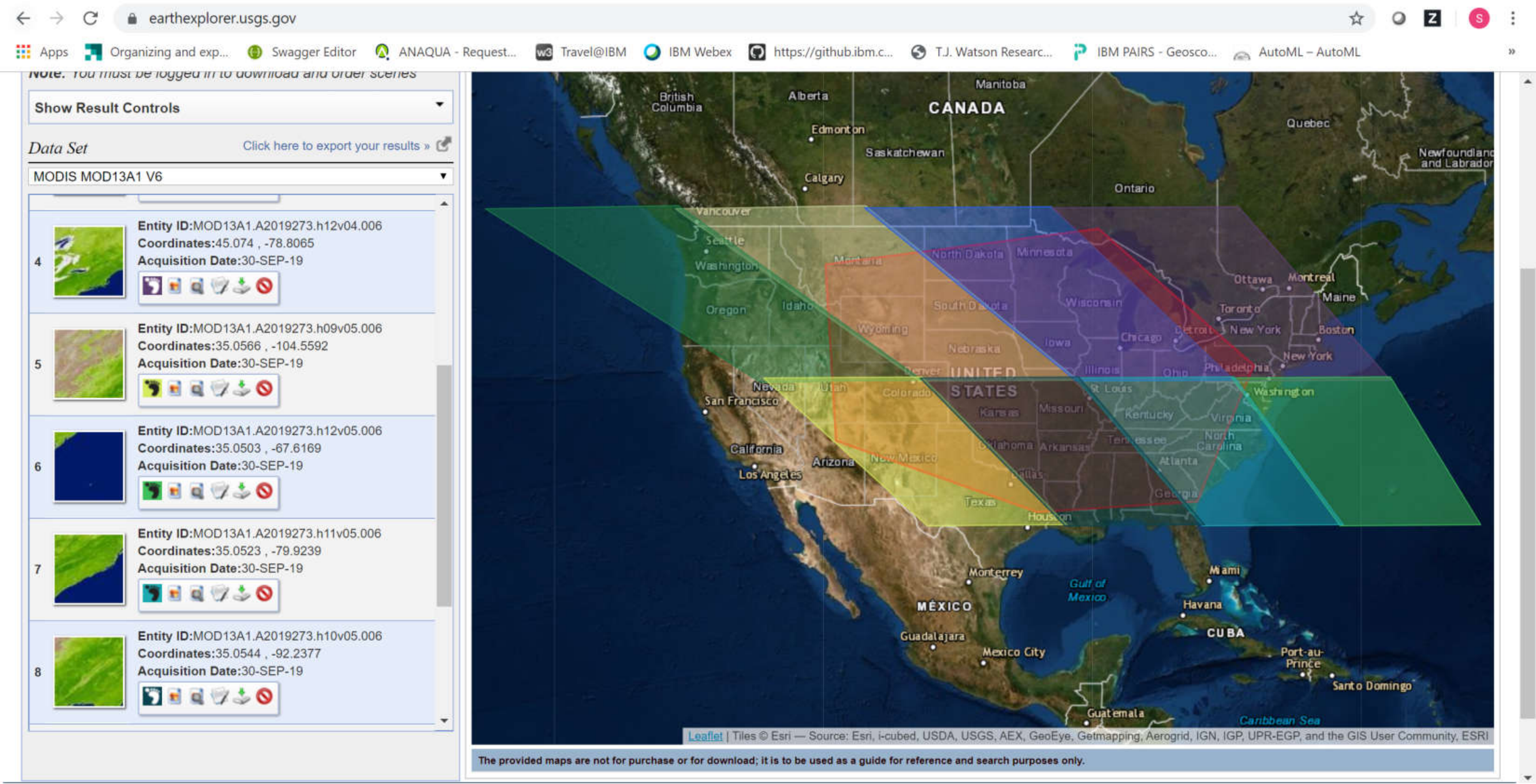
Multi-dimensional data

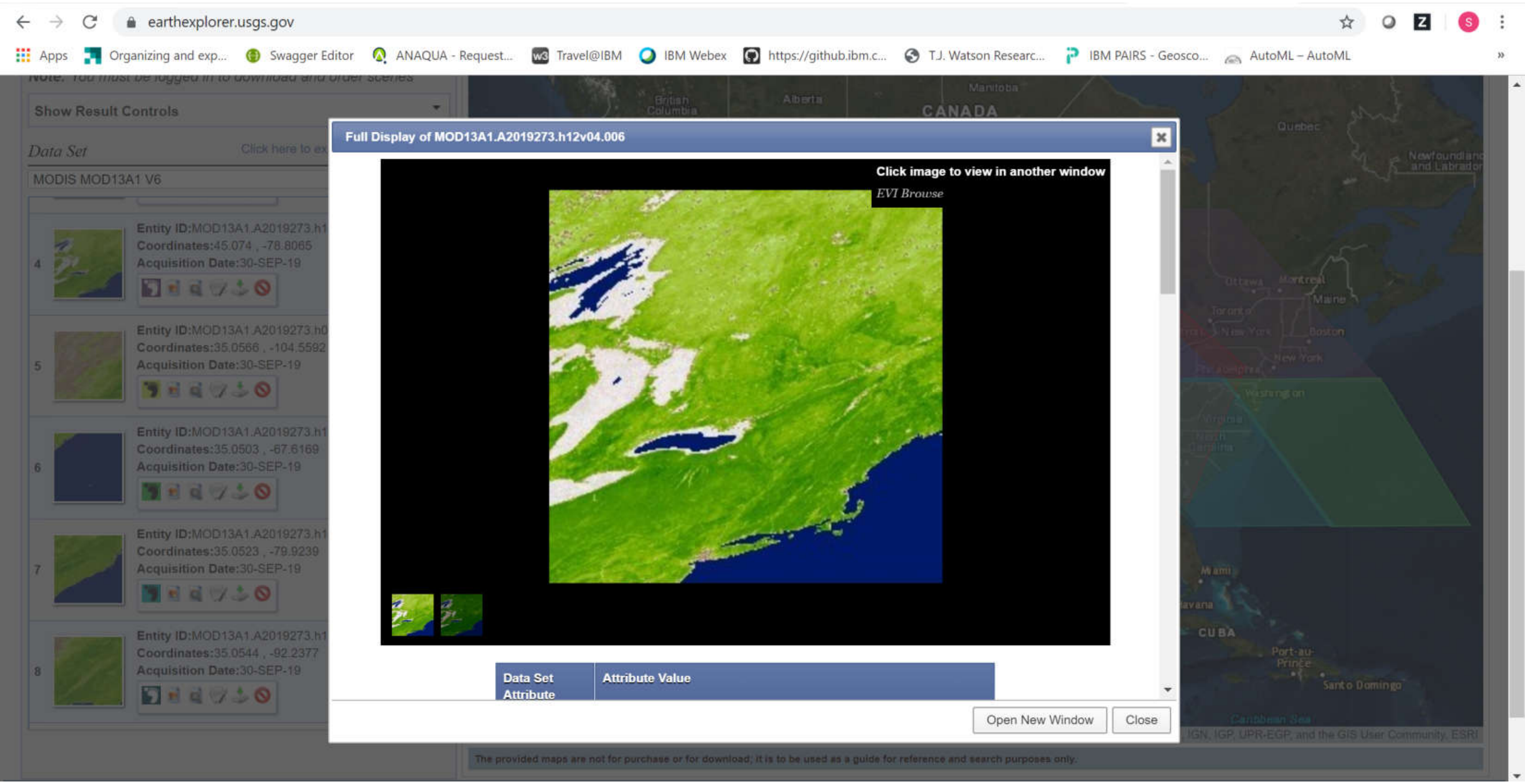


Key				Value
Row Key [128 bits]	Column		Version Timestamp	2^N x 2^N pixel super cell
	Column Family	Column Qualifier		
[4 bits reserved] + [4 bits Temporal Hash] + [4 bits Spatial Hash] + [52 bits Spatial] + [16 bits reserved] + [48 bits Temporal Key]	Data Layer	Additional Dimensions	Not used	Typical 32x32 pixels

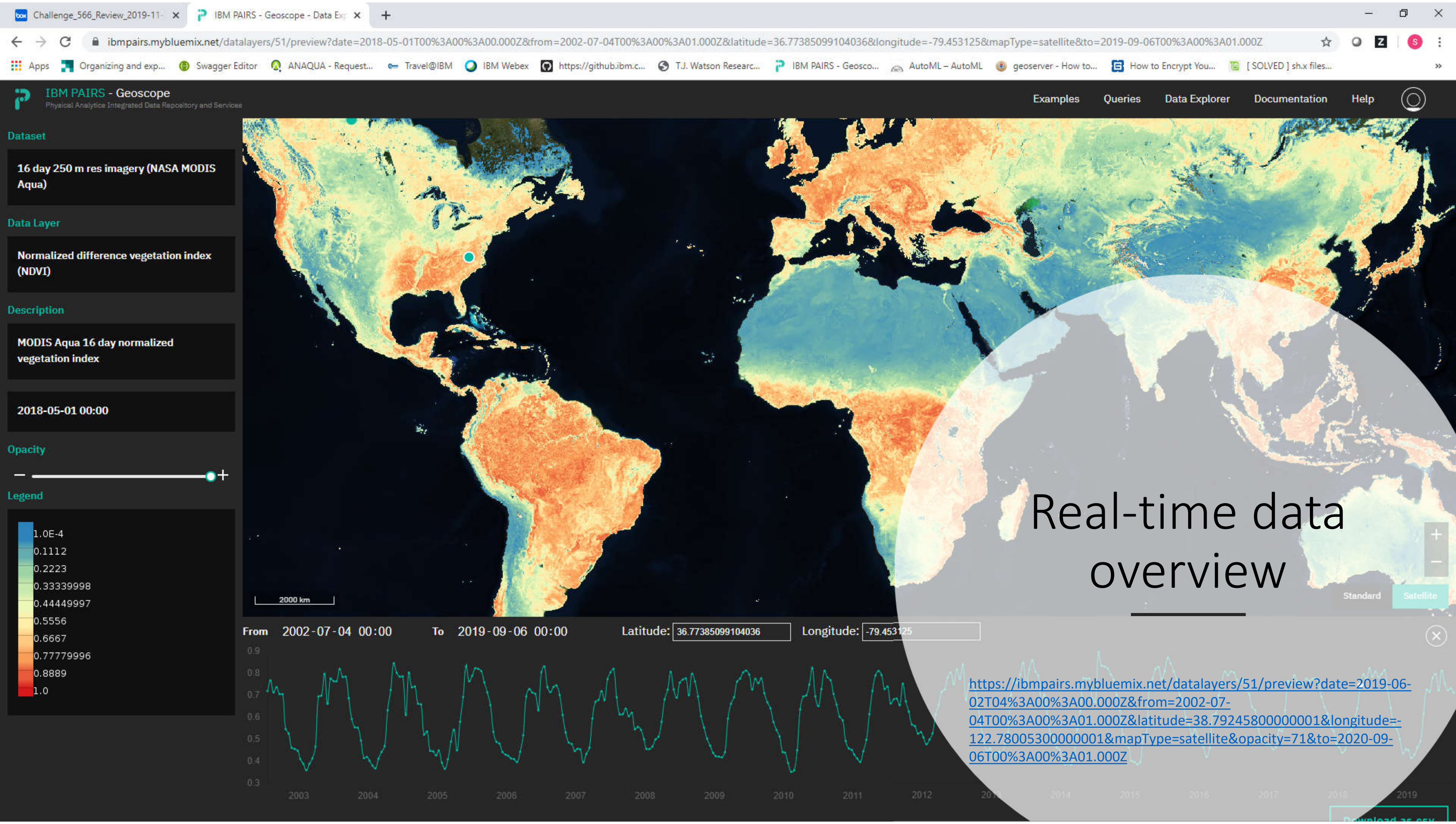
One dimensional key-value store

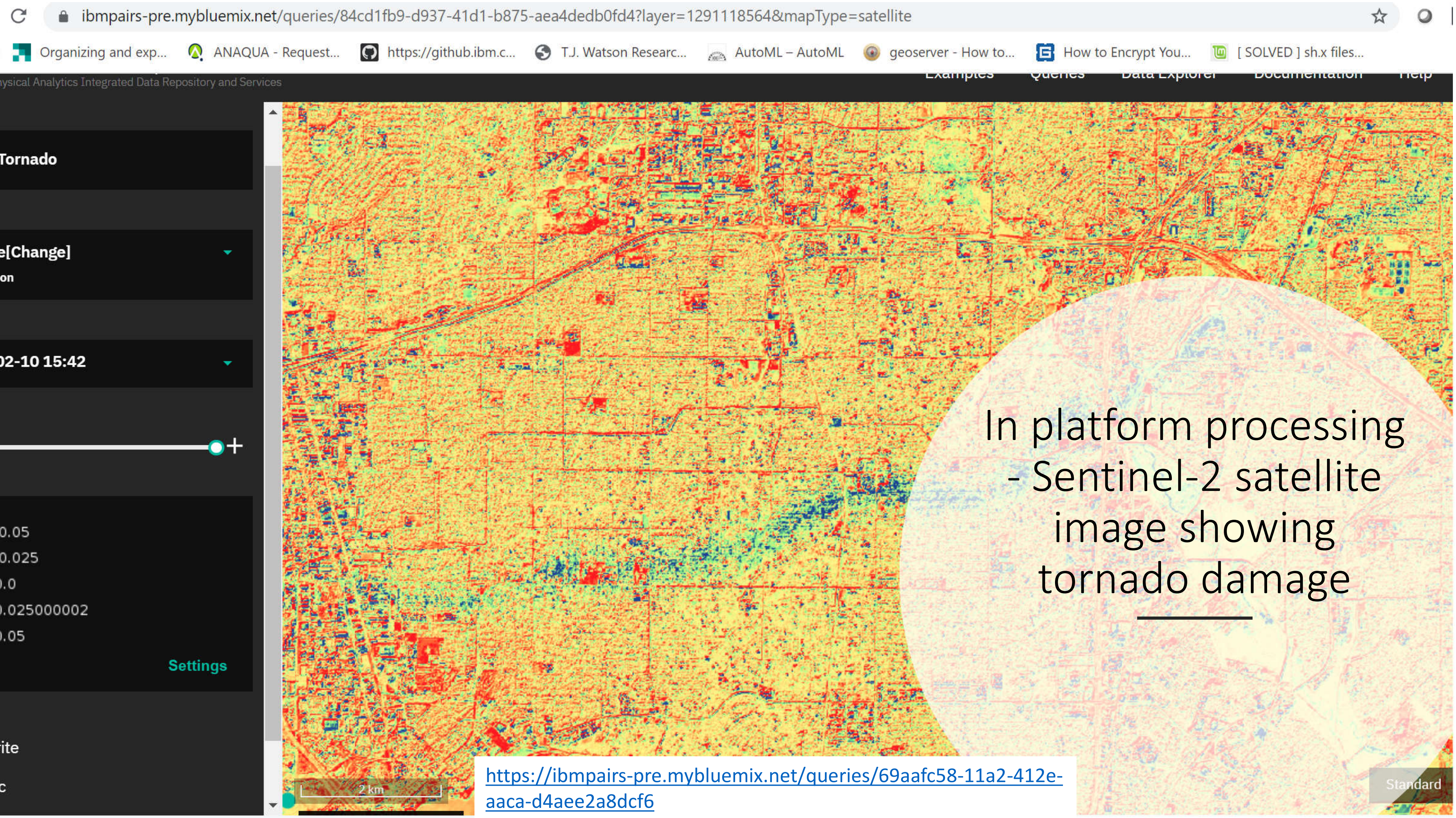

earthexplorer.usgs.gov
Apps
Organizing and exp...
Swagger Editor
ANAQUA - Request...
Travel@IBM
IBM Webex
https://github.ibm.c...
T.J. Watson Researc...
IBM PAIRS - Geosco...
AutoML – AutoML
EarthExplorer - Home
Page Expires In 1:59:50
Home
Login
Register
RSS
Feedback
Help
Search Criteria
Data Sets
Additional Criteria
Results
1. Enter Search Criteria
To narrow your search area: type in an address or place name, enter coordinates or click the map to define your search area (for advanced map tools, view the help documentation), and/or choose a date range.
Geocoder
KML/Shapefile Upload
Select a Geocoding Method
Address/Place
Address/Place
Show
Clear
Polygon
Circle
Predefined Area
Degree/Minute/Second
Decimal
1. Lat: 46° 50' 03" N, Lon: 112° 18' 51" W
2. Lat: 48° 50' 23" N, Lon: 089° 27' 33" W
3. Lat: 40° 45' 07" N, Lon: 076° 26' 58" W
4. Lat: 31° 32' 47" N, Lon: 081° 11' 47" W
5. Lat: 30° 47' 39" N, Lon: 094° 33' 28" W
Search Criteria Summary (Show)
Clear Search Criteria
(42° 35' 24" N, 158° 43' 38" W)
Options
Davis Strait
Hudson Bay
Gulf of Alaska
CANADA
UNITED KINGDOM
London
Paris
FRANCE
Madrid
SPAIN
Vancouver
Chicago
New York
San Franci
Los Angeles
UNITED STATES
Atlantic Ocean
ALGERIA
MEXICO
Mexico City
MALI
Caribbean Sea
Pacific Ocean
NIGERIA
Accra
Bogota
VENEZUELA
COLOMBIA
PERU
BRAZIL


earthexplorer.usgs.gov
Apps
Organizing and exp...
Swagger Editor
ANAQUA - Request...
Travel@IBM
IBM Webex
https://github.ibm.c...
T.J. Watson Researc...
IBM PAIRS - Geosco...
AutoML – AutoML
Show Result Controls
Data Set
Click here to export your results »
MODIS MOD13A1 V6
4
Entity ID:MOD13A1.A2019273.h12v04.006
Coordinates:45.074 , -78.8065
Acquisition Date:30-SEP-19
5
Entity ID:MOD13A1.A2019273.h09v05.006
Coordinates:35.0566 , -104.5592
Acquisition Date:30-SEP-19
6
Entity ID:MOD13A1.A2019273.h12v05.006
Coordinates:35.0503 , -67.6169
Acquisition Date:30-SEP-19
7
Entity ID:MOD13A1.A2019273.h11v05.006
Coordinates:35.0523 , -79.9239
Acquisition Date:30-SEP-19
8
Entity ID:MOD13A1.A2019273.h10v05.006
Coordinates:35.0544 , -92.2377
Acquisition Date:30-SEP-19
CANADA
UNITED STATES
MÉXICO
CUBA
Gulf of Mexico
Leaflet | Tiles © Esri — Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community, ESRI
The provided maps are not for purchase or for download; it is to be used as a guide for reference and search purposes only.



earthexplorer.usgs.gov
Show Result Controls
Data Set
MODIS MOD13A1 V6
Full Display of MOD13A1.A2019273.h12v04.006
Click image to view in another window
EVI Browse
Data Set Attribute
Attribute Value
Open New Window
Close
The provided maps are not for purchase or for download; it is to be used as a guide for reference and search purposes only.


IBM PAIRS - Geoscope
Physical Analytics Integrated Data Repository and Services
Examples
Queries
Data Explorer
Documentation
Help
Dataset
16 day 250 m res imagery (NASA MODIS Aqua)
Data Layer
Normalized difference vegetation index (NDVI)
Description
MODIS Aqua 16 day normalized vegetation index
2018-05-01 00:00
Opacity
Legend
1.0E-4
0.1112
0.2223
0.33339998
0.44449997
0.5556
0.6667
0.77779996
0.8889
1.0
2000 km
From 2002-07-04 00:00
To 2019-09-06 00:00
Latitude: 36.77385099104036
Longitude: -79.453125
Real-time data overview
https://ibmpairs.mybluemix.net/datalayers/51/preview?date=2019-06-02T04%3A00%3A00.000Z&from=2002-07-04T00%3A00%3A01.000Z&latitude=38.79245800000001&longitude=-122.78005300000001&mapType=satellite&opacity=71&to=2020-09-06T00%3A00%3A01.000Z



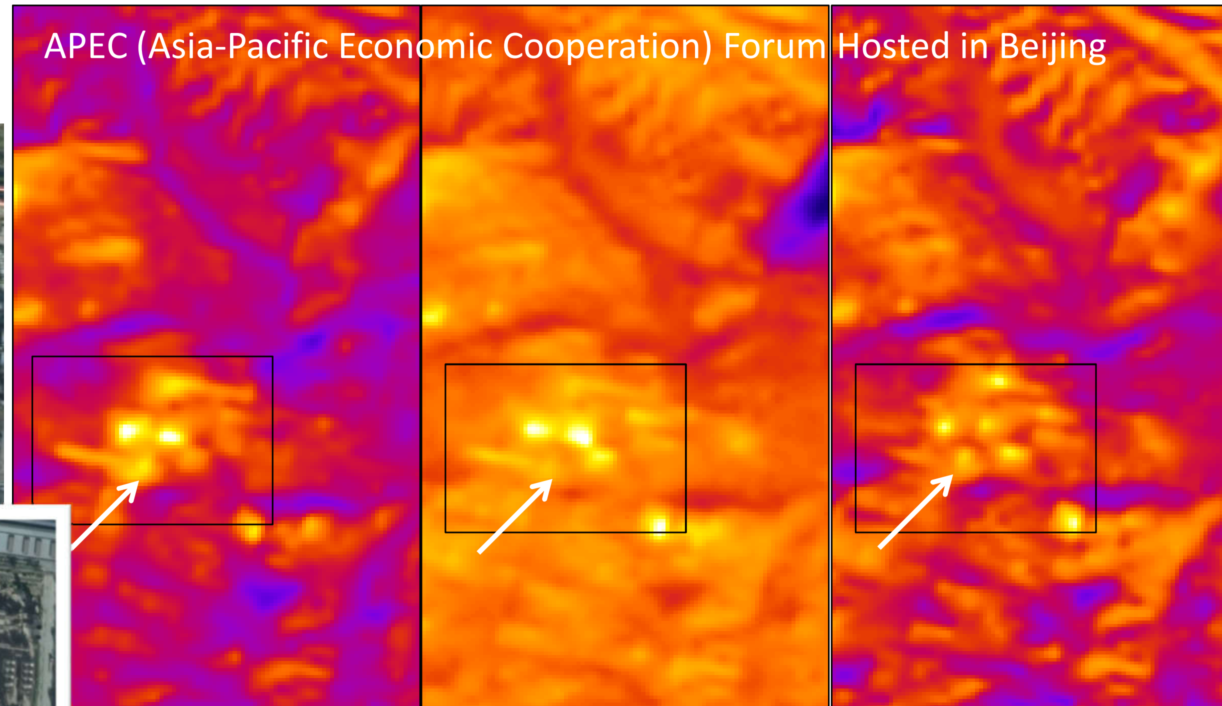
In platform processing - Sentinel-2 satellite image showing tornado damage

https://ibmpairs-pre.mybluemix.net/queries/69aafc58-11a2-412e-aaca-d4aee2a8dcf6

Thermal Infrared Monitoring of Steel Plants

Chengde Steel, Hebei, China

10/15/2014 11/16/2014 12/18/2014



Heat Generation priori to Statistical Correction

PAIRS enables user defined functions with a query (e.g., calculate radiative heat loss from LS8)

Command: curl -u username:password -X POST --header 'Content-Type: application/json' --header 'Accept: application/json' -d “example.json" 'https://pair.res.ibm.com/v2/query'

Example.json

```
{
  "layers": [ {  "alias": "LandSat_B10", "output": "false"        No output generated
              "id": “106,                                          LandSat 8 Level Thermal IR (band 10)
              "temporal": {"interval": [{"snapshot": "2014-11-16T12:00:00Z"}]}},
            {  "alias": "Surf_Temp“, "output": "false"
              "expression": "1321.08 / (math:log(774.89 / $LandSat_B10+1))",},   Calculate surface temperature
            {  "alias": "AmbientTemp", "output": "false"
              "id": "49257",                                       Ambient temperature from weather
              "temporal": {"interval": [{"snapshot": "2014-11-16T12:00:00Z"}]},},
            {  "alias": "RadiativeHeatLoss",
              "expression": "5.67e-8*( math:power($Surf_Temp,4) - 0.787*math:power($AmbientTemp,4) )" } ],   Calculate radiative heat loss
  "spatial": {"aoi": "43500", "type": "poly"},
  "temporal": {"interval": [{"snapshot": "2014-11-16T12:00:00Z"}]},
}
```

$$T_{Surf} = \frac{K_2}{\ln\left(\frac{K_1}{L_{\lambda 2}} + 1\right)}$$

$$Q_{rad} = \sigma \varepsilon_{LW} \left({T_{Surf}}^4 - \varepsilon_{LW}^{Air} {T_{amb}}^4\right)$$

Big Data PAIRS Query

PAIRS Queries

Example: Show me the croplands globally where the average summer temperature has risen more than 1.5 C in the last 40 years
(comparing 2009-2018 vs. 1979-1988)

2 different data layers

- *Global Crop Land*
- *ECMWF Weather Re-analysis*

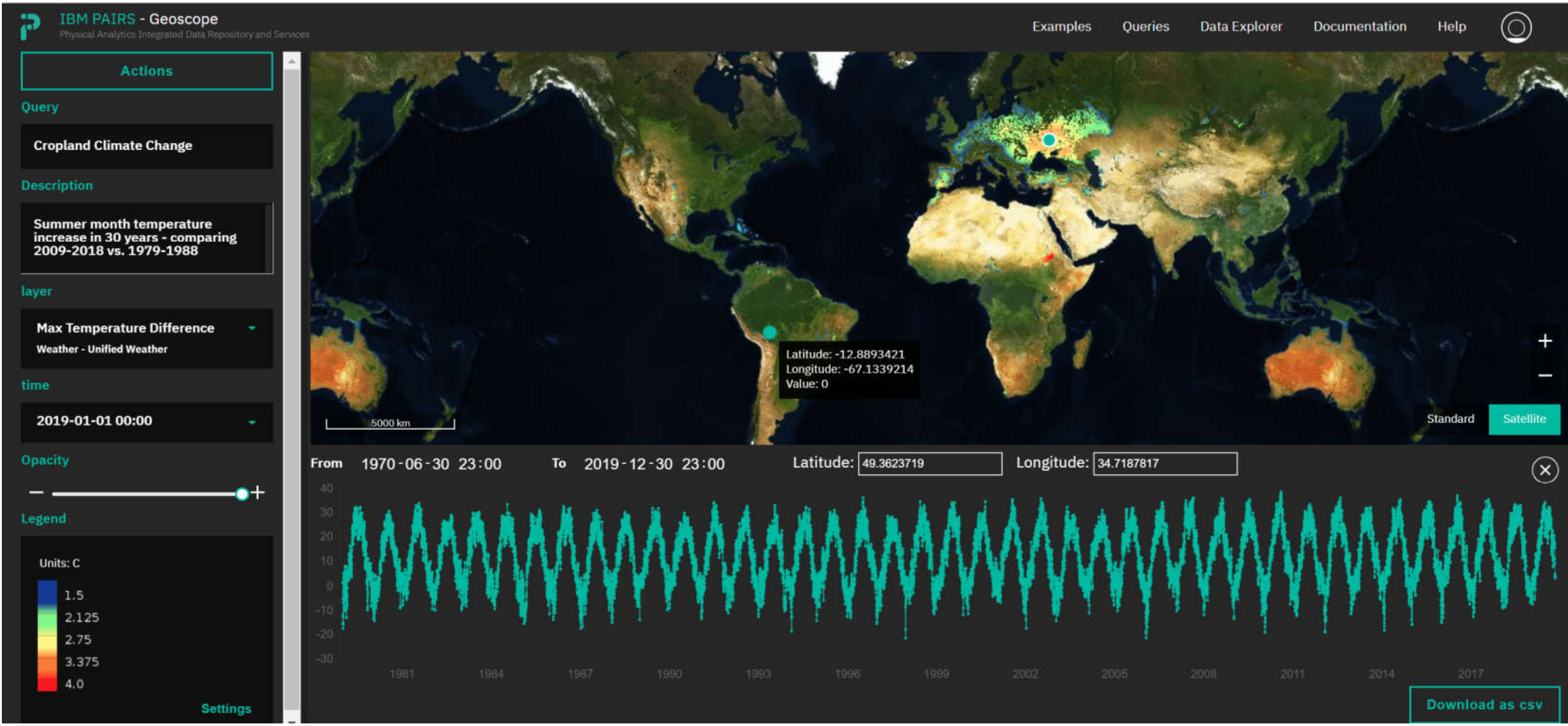
7361 timestamps
Global 250m grid

1 TB

```
query_payload = {
  'layers': [
    {'alias': 'Y2018','temporal': {'intervals': [{'start': '2018-06-01','end': '2018-08-31'}]},
      'id': 49188,'aggregation': 'Mean','output':'false'},
    {'alias': 'Y2017','temporal': {'intervals': [{'start': '2017-06-01','end': '2017-08-31'}]},
      'id': 49188,'aggregation': 'Mean','output':'false'},
    {'alias': 'Y2016','temporal': {'intervals': [{'start': '2016-06-01','end': '2016-08-31'}]},
      'id': 49188,'aggregation': 'Mean','output':'false'},
    {'alias': 'Y2015','temporal': {'intervals': [{'start': '2015-06-01','end': '2015-08-31'}]},
      'id': 49188,'aggregation': 'Mean','output':'false'},
    {'alias': 'Y2014','temporal': {'intervals': [{'start': '2014-06-01','end': '2014-08-31'}]},
      'id': 49188,'aggregation': 'Mean','output':'false'},
```

… ….

```
    {'alias': 'Y1980','temporal': {'intervals': [{'start': '1980-06-01','end': '1980-08-31'}]},
      'id': 49188,'aggregation': 'Mean','output':'false'},
    {'alias': 'Y1979','temporal': {'intervals': [{'start': '1979-06-01','end': '1979-08-31'}]},
      'id': 49188,'aggregation': 'Mean','output':'false'},
    {
      'alias': 'TempDiff',
      'expression': '($Y2018 + $Y2017 + $Y2016 + $Y2015 + $Y2014 + $Y2013 + $Y2012 + $Y2011 + $Y2010 + $Y2009)/10 \
                     - ($Y1988 + $Y1987 + $Y1986 + $Y1985 + $Y1984 + $Y1983 + $Y1982 + $Y1981 + $Y1980 + $Y1979)/10',
      'filter':'GT 1.5'
    },
    {'alias': 'crop_dominance','temporal': {'intervals': [{'snapshot': '2017-01-01'}]},'aggregation': 'Mean',
      'id': 49546, 'filter':'GT 0.5'
    }
],
  'temporal': {
    'intervals': [
      {
        'start': '1979-06-01',  'end': '2018-08-31'
      }
    ]
  },
  'name': 'Aggregated North Semisphere Filter 10 year' ,
  'spatial': {
    'type': 'square',
    'coordinates': [
      0,-179, 80, 179
    ] #, 'aggregation': {'aoi': [23]}
  }
}
```



Example: Show me the croplands globally where the average summer temperature has risen more than 1.5 C in the last 40 years
(comparing 2009-2018 vs. 1979-1988)

Big Data PAIRS Query

Example: Show me the croplands globally where the average summer temperature has risen more than 1.5 C in the last 40 years
(comparing 2009-2018 vs. 1979-1988)

2 different data layers

- *Global Crop Land*
- *ECMWF Weather Re-analysis*

7361 timestamps
Global 250m grid

1 TB

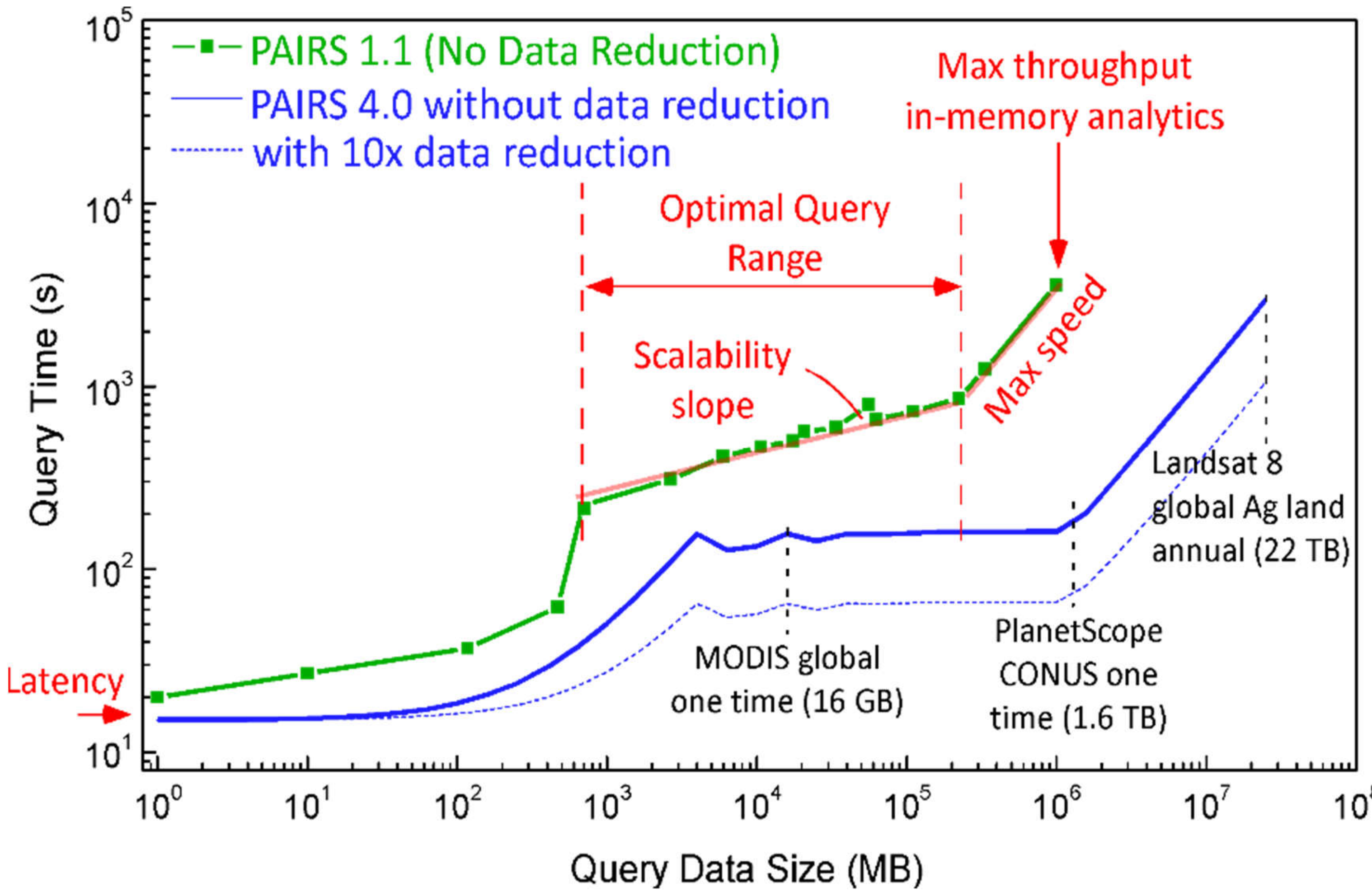
Result:

0.93 TB /61 s = 15 GB/s
= 1.3 PB/day

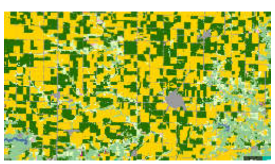
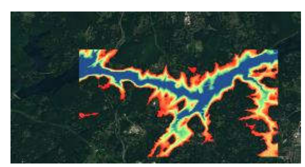
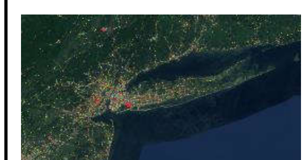
We can run >5 queries without performance degradation

= 6.5 PB/day

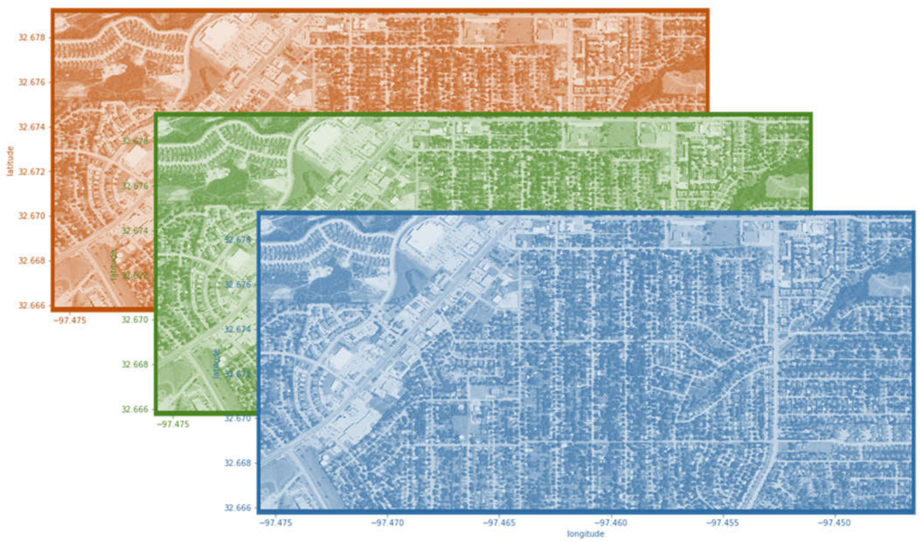
How to think about raster query performance?



PAIRS industry applications are plentiful

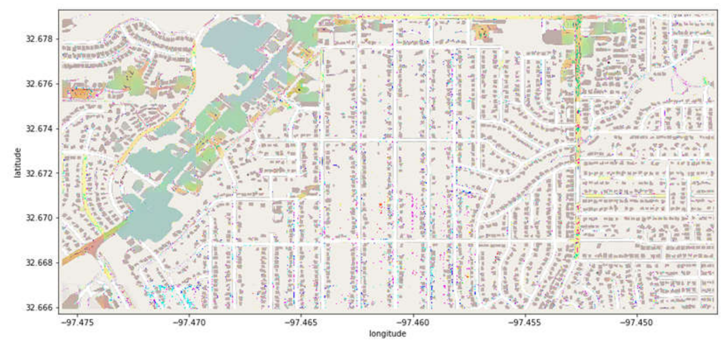
Industry	Finance ✓	Utility ✓	Agriculture ✓	Insurance ✓	Retail	Public ✓
Example	Commodity trading	Vegetation management	Decision support for agriculture	Develop risk models (Flood, Fire)	Supply change	Disaster response
Example Queries / Questions	How much crop is planted? How much corn will be produced in Argentina?	Where do trees infringe on utility assets? When to schedule tree trimming?	What is the best crop to plant? When to apply fertilizers, pesticides?	Where are my assets at risk? At what time of the year is the risk the highest ?	Where, when and what to ship? Where should I promote a product?	Who is being impacted by a hurricane? What is the best emergency response?
Data Layers	Weather, land class, soil, satellite	Weather, satellite, power line data	Weather, land class, soil, satellites	Climate, vegetation, traffic, census	Weather, socio-economic data, store locations	Weather, climate, satellites, map, socio-economic data
Example	**Early Crop Recognition** 	**Tree management** 	**Precision Agriculture** 	**Flooding Risk** 	**Optimal store locations** 	**Emergency Response** 

Spatial Analytics – Raster-to-Raster



… …

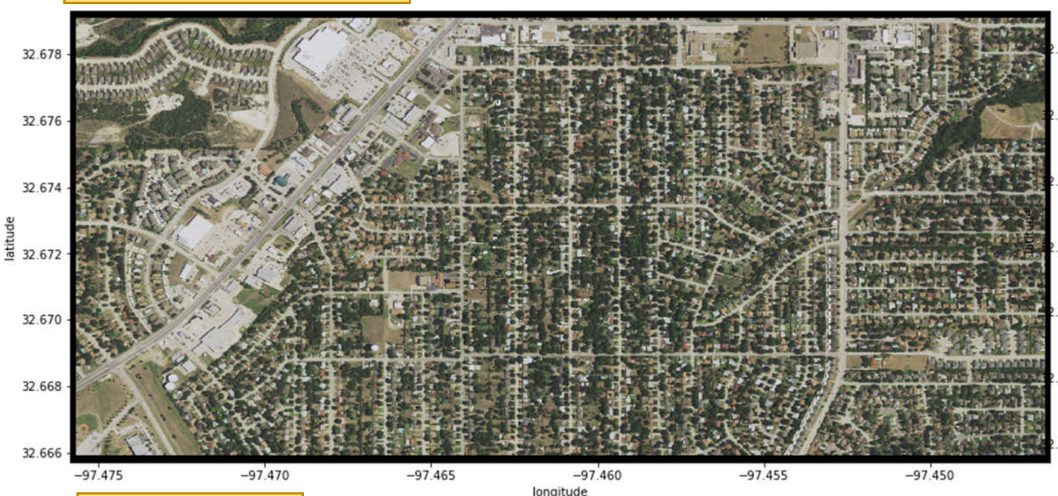
R/G/B/(NIR) of Satellite Images



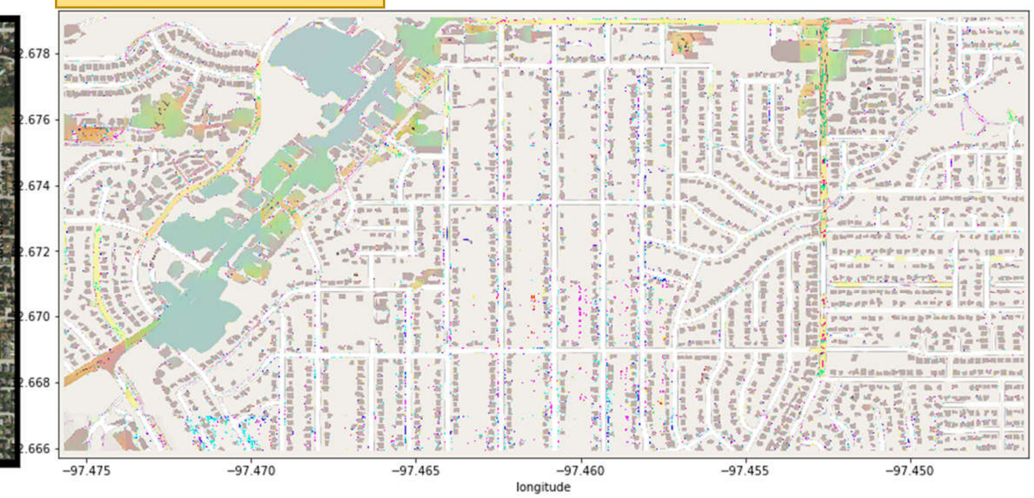
Map features:
House, Road, Forestry …

Road and House Detection Using High Resolution Imagery and OpenStreetMap

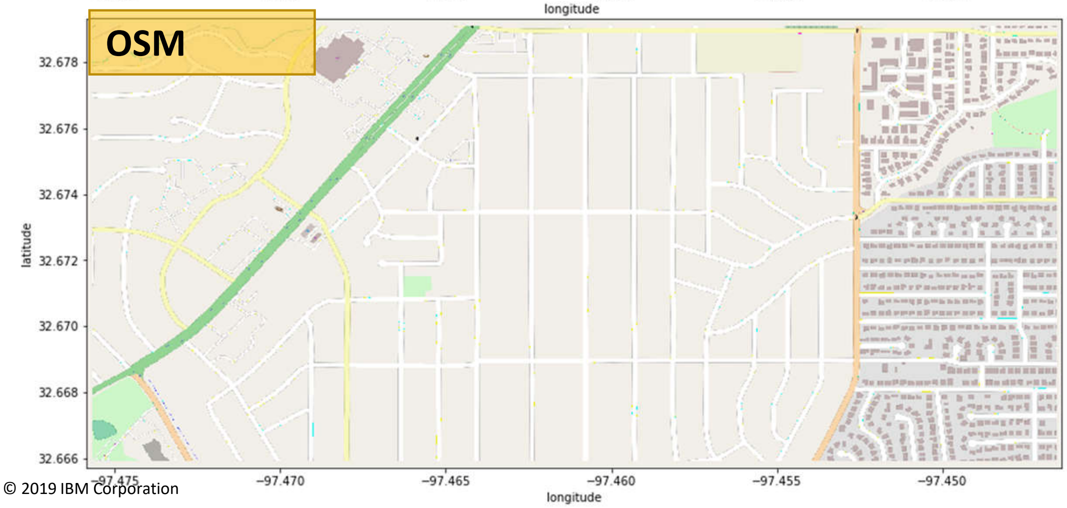
NAIP imagery



Generated MAP



OSM



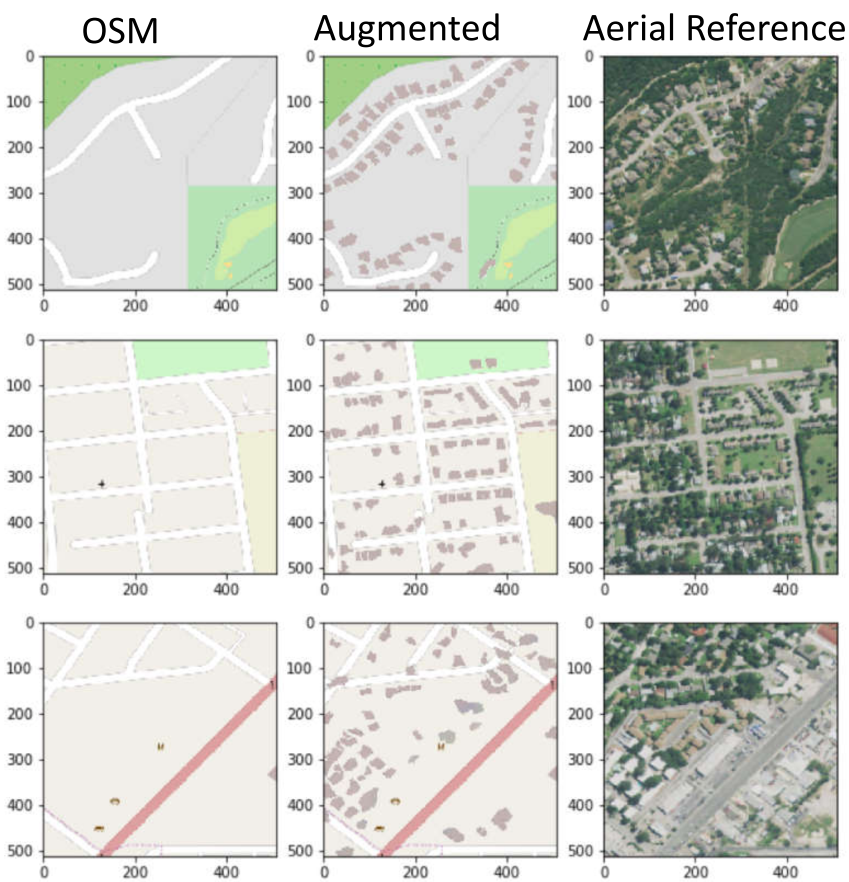
Engineering Challenge

- CONUS high res. imagery ~ 100 TB. How to efficiently parallelize deep learning on large number of GPUs

Algorithmic Challenge

- Input data inaccuracy (e.g. Open street map's road is shifted from where it should be). How to mitigate?

Data augmentation enables ML on “crappy” OSM data to match ML on accurately labeled SpaceNet



Before Augmentation

Model	Precision	Recall	F1
FW-CycleGAN	0.641	**0.740**	0.687
U-Net	0.816	0.732	0.772

After Augmentation

Model	Precision	Recall	F1
FW-CycleGAN	0.844	0.811	0.828
U-Net	0.950	0.811	**0.875**

SpaceNet Winner F1 0.88

Change Detection (Out of sample results)

2010

2016

Physical Analytics @ IBM T. J. Watson Research Center


Conrad Albrecht
(Physical modeling,
Heidelberg PhD)

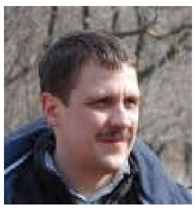

Norman Bobroff
(Big Data,
Caltech PhD)


Bruce Elmegreen
(Astronomy,
Princeton PhD)


Marcus Freitag
(Precision Ag,
U.Penn, PhD)


Hendrik Hamann
(Physical Analytics
Goettingen, PhD)


Ildar Khabibrakhmanov,
(Geospatial software, Moscow
Inst. of Phys.Tech., PhD)


Levente Klein
(Physical Modeling,
U. Utah PhD)


Theodore van Kessel
(Oil and Gas)


Siyuan Lu
(Physical Analytics
USC, PhD)


Fernando Marianno
(Software Architect)


Michael Schappert
(Computing System)

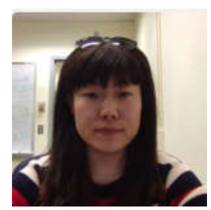

Johannes Schmude
(Machine Learning,
Swanson Univ. PhD)


Xiaoyan Shao
(Machine Learning,
John Hopkins PhD)


Carlo Siebenschuh
(Karlsruhe Institute
of Technology)


Fateh Tipu
(Geospatial software)


Dharmesh Vadgama
(Software developer)


Rui Zhang
(DL, CMU, PhD)


Wang Zhou, (DL,
Northwestern, PhD)

PAIRS Resources

PAIRS was commercialized by IBM TWC Business Unit in Nov 2018.

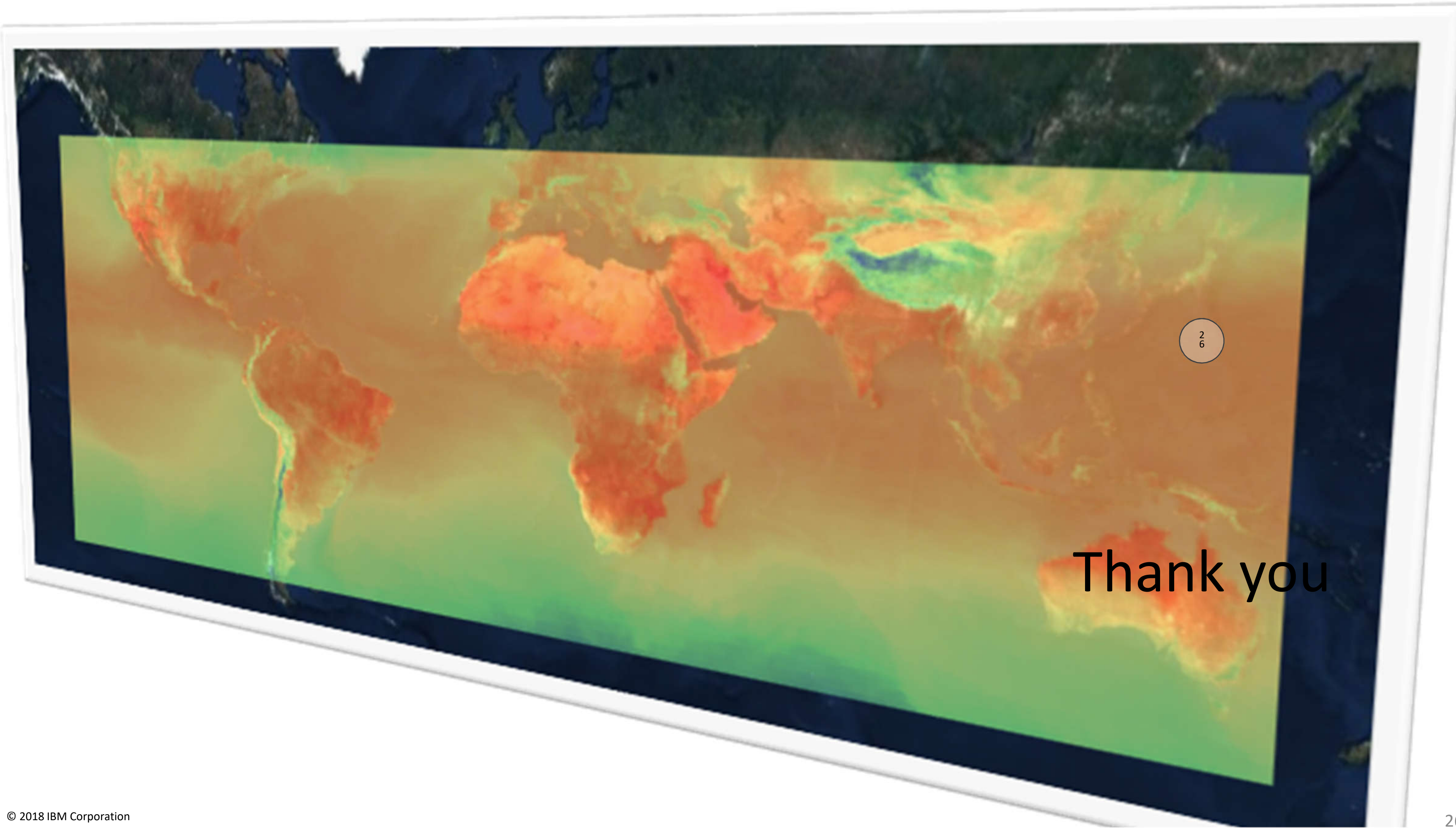
Freemium GUI: https://ibmpairs.mybluemix.net

API entry point: https://pairs.res.ibm.com

API tutorial and reference: https://pairs.res.ibm.com/tutorial/

Data documentation: https://ibmpairs.mybluemix.net/data-explorer

Python SDK (open source): https://github.com/ibm/ibmpairs



Thank you